DOCUMENT RESUME

ED 472 166 TM 034 720

AUTHOR	Yang, Yu Nu
TITLE	Testing the Stability of Experts' Opinions between Successive Rounds of Delphi Studies.
PUB DATE	2003-04-00
NOTE	16p.; Paper prepared for the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Delphi Technique; Nonparametric Statistics; *Reliability; Simulation
IDENTIFIERS	*Experts; McNemar Test of Equality of Correlated Proportion; Stopping Rules

ABSTRACT

The Delphi method is a means of structuring group communication process so that a group of experts can gather information or forecast future problems effectively. A primary objective of a Delphi study is to obtain consensual and consistent opinions from a group of experts in two or more successive rounds on a given research subject. Consensus and consistency are presumed to have been reached when a stopping criterion used for determining a consensus has been met. This study examined two nonparametric statistical methods: the McNemar change test (Q. McNemar, 1969) and the Binomial test for setting stopping rules in the context of Delphi studies. Simulation results indicate that the McNemar change test could be robust even if the correction for continuity was not made, and the McNemar test was not as conservative as the Binomial test. Discussions on how to apply both tests to Delphi studies are included. (Contains 4 tables and 19 references.) (Author/SLD)

Testing the Stability of Experts' Opinions between Successive Rounds of Delphi Studies

by

Yu Nu Yang

National Chia-Yi University

Paper presented at the 2003 annual meeting of the American Educational Research Association, Chicago, IL.

Testing the Stability of Experts' Opinions between Successive Rounds of Delphi Studies

Abstract

The Delphi method is a means of structuring a group communication process so that a group of experts can gather information or forecast future problems effectively. A primary objective of a Delphi study is to obtain consensual and consistent opinions from a group of experts in two or more successive rounds on a given research subject. Consensus and consistency are presumed to have been reached when a stopping criterion used for determining a consensus has been met. This paper examined two nonparametric statistical methods (e.g., The McNemar change test and the Binominal test) for setting stopping rules in the context of Delphi studies. Results indicates that: (a) The McNemar change test could be robust even though the correction for continuity were not made and (b) The McNemar test is not as conservative as the Binominal test. Discussions on how to apply both tests to Delphi studies were included.

Key Words: Delphi Studies, McNemar change test, Binominal test, Chi-square test

Paper presented at the 2003 annual meeting of the American Educational Research Association, Chicago, IL.

Testing the Stability of Experts' Opinions between Successive Rounds of Delphi Studies

I. Introduction

The Delphi method is a means of structuring a group communication process so that a group of experts can gather information or forecast future problems effectively (Linstone & Turoff, 1975). Once a panel of experts is identified and selected, they are asked to render opinions and judgments on an open-ended and structured questionnaire. Qualitative analyses (e.g., content analysis, Tafoya, 1986) are performed on the panelists' responses to the questionnaire and a Likert-type questionnaire is then developed using the results of these analyses. The next phase asks the same panelists to rank the items based on a Likert scale measuring their importance, desirability or feasibility. Following that, several rounds may be needed until a consensus on each item has been achieved based on the results of the quantitative analysis (discussed later). It is apparent that the Delphi method combines the qualitative analysis with the quantitative analysis in its research design. This methodology has been employed in a variety of areas such as health care (Niero & Robertson, 1996; Toohey & Shireffs, 1980), communication and public relations (Blamphin, 1990; Preble, 1983), and art education research (e.g., Ferris, 1998; Hendricks, 1995; Lucas1986; Wang, 1992; Yang, 2000).

A primary objective of a Delphi study is to obtain consensual and consistent opinions from a group of experts in two or more successive rounds on a given research subject (Dajani, Sincoff & Talley, 1979). Consensus and consistency are presumed to have been reached when a stopping criterion used for determining a consensus has been met. However, problems arise from the use of stopping criteria since they are rather arbitrary and subjective. For instance, a change of less than 15% between two successive-round average scores for an item were considered to be a proper stopping rule for this item (Franchack, Desy and Norton, 1984).

If we anticipate that the public is more amenable to accepting the consensual conclusions resulting from this methodology, we should rely on statistical techniques rather than arbitrary criteria (refer to Dajani, Sincoff & Talley, 1979) for determining at which point the rounds should be discontinued. Dajani, Sincoff and Talley (1979) have reviewed several methods that have been used in Delphi studies and have proposed the chi-square test for serving this purpose. The chi-square test introduced in that study is suitable for testing whether two independent samples have any significant different responses to the research topic of interest. Most Delphi studies, however, use the same group of samples in two successive rounds and make them answer almost identical questions repeatedly. It appears that other statistical tests are appropriate to this Delphi-study context and are introduced and examined in this study.

This paper further reviews various statistical methods used by or suitable for Delphi studies for setting stopping rules. It begins with descriptions of several parametric statistics that have been used in Delphi studies, followed by illustrations of statistical methods that are more appropriate in the context of Delphi studies. Evaluation of selected methods using simulated data is presented in the fourth section, and the conclusion is presented in the final section of this paper.

II. Parametric Statistical Methods

1. The Coefficient of Variation.

The coefficient of variation (CV) is the ratio of the standard deviation of an item's rating score to its corresponding mean among panelists. This coefficient, unlike the standard deviation, is not affected by the unit of measurement. For example, it would be the same if the incomes of a group of people were measured in hundreds of dollars or tens of hundreds of dollars. English & Kernan (1976) has used the CV to determine the stopping rule. If the magnitude of CV for one item is relatively too large (e.g., >0.8), the corresponding statement may need to be modified and an additional round(s) becomes necessary. In contrast, an additional round is not needed if CV is less than or equal to 0.5. The determination of this criterion, however, remains an arbitrary matter.

According to Dajani (1979), stability is considered reached when the absolute value of the difference in CV between two rounds reaches a minor difference. Otherwise, more rounds are required. Ideally, an item's CV value obtained from the latter round should be smaller than that obtained in an earlier round.

2. The Degree of Association

Another alternative stability measurement is the level of agreement between two round ratings among experts on the same item. The Pearson correlation coefficient was used for this measure. If a statement's correlation coefficient varies significantly from zero and is very high, the experts' ratings on this item are stable and are less fluctuated. Otherwise, this item's ratings vary significantly between two rounds.

3. F-test in Comparing Two Variances

The observed F value is defined by the ratio of the variances of item scores among panelists between the two successive rounds. The questionnaire item will be dropped from further rounds when no significant difference in the F test is identified. Only those questionnaire items in which a significant between-round difference in variance is found are retained in a subsequent round. This method has been suggested by Jolson and Rossow (1971), however, the problem with this method is that the assumptions made for the F-test may be seriously violated when using data that is collected from the Delphi rounds.

III. Nonparametric Statistical Methods

1. McNemar Change Test

Suppose a panel of 24 Art-education experts were asked in two successive rounds in a Delphi study about their attitudes toward the question: Can Web-based technologies be integrated into the design and development of art teacher education curricula by using the model of Discipline-Based Art Education (DBAE) as an integrated method of learning?

The resulting data are presented in Table 1. It should be noted that the order of the data entries in Table 1 is important. The cells marked A and D are those indicating a change in response from the previous round to the current round. The cells marked B and C are those indicating no change between two successive rounds.

Table 1. A sample of data on the attitude toward the web-based technology in art teacher education curriculum

		Round 2		Subtotal
		Disagree	Agree	
Round 3	Agree	A (9)	B (8)	17
	Disagree	C (2)	D (5)	7
	Subtotal	11	13	24

In the above example, we are interested only in the cells of the contingency table that reflect a change of opinion about the question; namely, cells A and D. A specific null hypothesis is that there will be an equal number of changes in both directions in the target population (e.g., art-education experts). This implies that the expected frequencies in cell A will be equal to the expected frequencies in cell D. If the null hypothesis is retained, we might make an inference that the experts' opinions on the question were stable or consistent between the two successive rounds. If the null hypothesis is rejected (or the alternative hypothesis is retained), we might infer that the experts' opinions on the question were inconsistent between the two successive rounds.

Under the above scenario, the McNemar change test is appropriate. The McNemar test statistic (McNemar, 1969) is computed as:

$$\text{McNemar change statistic} = \frac{(A-D)^2}{A+D} \tag{1}$$

The approximate sampling distribution of chi-square calculated from Equation 1 becomes more precise if a correction for continuity is made by using the following equation (refer to Siegel & Castellan , 1988):

$$\frac{(|A-D|-1)^2}{A+D} \tag{2}$$

Since there are only two cells (A and D) under consideration, once the expected frequencies for one of the two cells are known, the expected frequencies of the other are thereby determined. Thus, there is 1 degree of freedom associated with the McNemar change statistic. The critical value of chi-square for 1 degree of freedom at type-I error =.05 is 3.841. For the data in Table 1, the McNemar change statistic was computed as:

$$\text{McNemar Chi-square statistic} = \frac{(|A-D|-1)^2}{A+D} = \frac{(|9-5|-1)^2}{9+5} = \frac{9}{14} = 0.64$$

Since the computed McNemar chi-square value, 0.64, was larger than the critical value (3.841), the null hypothesis was not rejected, and the conclusion is that the experts' attitudes toward the example question have not significantly changed in one direction to the other.

In an earlier note, Dajani et al. (1979) applied a chi-square test to Delphi studies. That chi-square test was supposed to be used for two independent samples. In order to compute the chi-square statistic for the condition of two independent samples, data in Table 1 were modified for creating Table 2, in which each round has a sample size of 24 (note: here 24 samples in Rounds 2 and 3 were treated as two independent groups).

Table 2. Data modified from table 1 for computing the chi-square statistic for two independent samples

		Response		Subtotal
		Disagree	Agree	
	Round 2	11	13	24
	Round 3	7	17	24
Subtotal		18	30	48

The chi-square value calculated from the data in Table 2 is 1.42 that is different from the one calculated from the McNemar change test. This difference reminds researchers not to apply the chi-square test that is used for two independent samples to the data collected from the Delphi studies.

However, when the expected frequency for the McNemar change test, (A+D)/2, is very small (say less than 5), the approximation for chi-square value in Equation 2 may be poor. Under this circumstance, the binominal test is preferred (Siegel & Castellan, 1988). The binominal test is illustrated next.

2.The Binomial Test

There are many types of data that are conceived as consisting of only two classes. Examples of such classes are: member and non-member, married and single. For the sample size, N, the probability of obtaining k objects in one category along with N- k objects in the other category is given by (refer to Siegel & Castellan, 1988):

$$P[Y = k] = \binom{N}{k} P^{k} q^{N-k} \qquad k=0, 1,\ldots, N \tag{3}$$

where

p= the proportion of observations expected where X=1,

q= the proportion of observations expected where X=0, and

$\binom{N}{k}$ is the number of combinations of size k that can be constructed from N distinct objects.

Table 3. Data for dependent samples in the attitudes toward the web-based technology in art teacher education curriculum

		Round 2		Subtotal
		Disagree	Agree	
Round 3	Agree	A (5)	B (17)	22
	Disagree	C (0)	D (2)	2
	Subtotal	5	19	24

When applying the binominal test to the Delphi studies, three key components are essential. They are: (a). The population parameters for p and q should be set at 0.5, (b). The null hypothesis would be that the sample of N (note: N=A+D) cases came from a binominal population, and (c) The value of k is the smaller of the two observed frequencies, either A or D.

For the data in Table 3, N=7 and k=2. Using the probability table (usually given in statistical textbooks) for the binominal test, the probability associated with values as small as (or smaller than) observed values of 2 from the 7 trials is .227. When doubling this probability, it yields the probability associated with the two-tailed change test, which, for this example, is .454. This result implies that experts did not significantly change their opinions from Round 2 to Round 3.

IV. Evaluation of Selected Methods

A. Simulated Data

The dataset used for chi-square test was simulated as similar to real data as possible. The number of responses in the contingency tables for thirty-two items is presented in Table 4, in which sample size was set to 24 for all items. The labels of A, B, C and D represent the cells of the fourfold table defined as Table 1. For instance, A =1 in Item 1 represents one expert who disagreed with this research topic at Round Two stage, but agreed with this research topic at Round Three stage.

Because the total number of "changes," A +D, will be used for determining an appropriate statistical test (either McNemar change test or Binomial test), the value of A+D was first determined. Following that, the combination of A along with D values was determined based on two principles. One is that the values of A and D are expected to be similar to the data obtained from a real Delphi study. The other is that the combination of A and D will be more likely to result in varying chi-square statistics so that a greater variety of the observed chi-square values can be obtained and tested.

B. Results and Discussions

The McNemar change test without correction (called M1), the McNemar test with correction (called M2), the Binomial Test (called M3) , and the chi-square test (called M4) for the condition of two independent samples were used. The statistical test regarding whether experts have reached an agreed conclusion on each item was conducted by each of the four statistical methods. The labels of N and Y in Table 4 represents "No change of opinions" as well as "Change of opinions". For instance, experts did not change their opinions for Item 1 based on M1, M2, M3 or M4 statistical test. The letters of Y and N with shaded gray color are the presumed correct decisions being made. For example, the Binomial test is suitable for Item 6 with A+D=8 and the decision of "No Changing Opinion" made by the Binomial Test is presumed correct. The letter of Y or N with star sign indicates that the decision made by the corresponding statistical method could be wrong.

Table 4: Simulated Dataset and Significant Tests by Four Statistical Methods

Items	A	B	C	D	A+D	M1	M2	M3	M4
1	1	9	9	5	6	N	N	N	N
2	2	12	6	4	6	N	N	N	N
3	1	9	8	6	7	N	N	N	N
4	2	10	7	5	7	N	N	N	N
5	3	3	14	4	7	N	N	N	N
6	1	8	8	7	8	Y*	Y*	N	N
7	2	2	14	6	8	N	N	N	N
8	3	9	7	5	8	N	N	N	N
9	1	10	5	8	9	Y	Y	Y	Y
10	2	7	8	7	9	N	N	N	N
11	3	9	6	6	9	N	N	N	N
12	4	1	14	5	9	N	N	N	N
13	1	6	8	9	10	Y	Y	Y	Y
14	2	2	12	8	10	N	N	N	N
15	3	7	7	7	10	N	N	N	N
16	4	10	4	6	10	N	N	N	N
17	1	12	1	10	11	Y	Y	Y	Y
18	2	8	5	9	11	Y	Y	N*	Y
19	3	4	9	8	11	N	N	N	N
20	4	10	3	7	11	N	N	N	N
21	5	7	6	6	11	N	N	N	N
22	1	6	6	11	12	Y	Y	Y	Y
23	2	3	9	10	12	Y	Y	Y	Y
24	3	8	4	9	12	N	N	N	N
25	4	3	9	8	12	N	N	N	N
26	5	4	8	7	12	N	N	N	N
27	1	1	10	12	13	Y	Y	Y	Y
28	2	5	6	11	13	Y	Y	Y	Y
29	3	7	4	10	13	N	N	N	Y*
30	4	3	8	9	13	N	N	N	N
31	5	6	5	8	13	N	N	N	N
32	6	9	2	7	13	N	N	N	N

M1: McNemar change test without correction, M2: McNemar change test with correction
M3: Binomial Test, M4: Chi-square test for the condition of two independent samples

The results presented in Table 4 show that McNemar test with correction vs. without correction made no difference across all items. This implies that the McNemar change test could be rather robust even though the correction for continuity was not made. It is noted that the simulated data did not include all possible combinations of the values for the cells and A, B, C, and D. Hence, both methods could yield different results for some types of contingency tables. The McNemar with correction is still preferred.

Under the condition of A+D >10, choosing the McNemar test with correction is a legitimate method for decision-making. However, if the Binomial test was used instead of McNemar test with correction, Item 18 was misidentified as "No change" although the decision of "change opinion" was presumed appropriate.

Choosing the Binomial test is a legitimate method for decision-making for the condition of A+D <=10. However, if the McNemar test with correction was used instead of the Binomial test, Item 6 was misidentified as "changing opinion" although the decision of "No change" was presumed correct.

Comparing the results between the McNemar test with correction and the Binomial test, the findings suggested that the Binomial test is more conservative than the McNemar change test. Hence, if we apply the Binomial test to all items regardless of the condition of the total number of changes, we may conclude that some items (e.g., item 18) that should be identified as "Changing Opinion" are identified as "No change". On the other hand, if we apply the McNemar test with correction to all items regardless the total number of changes, we may conclude that some items (e.g., item 6) that should be identified as "No change Opinion" are identified as "Changing Opinion".

As noted previously, the chi-square test using two independent samples is not a legitimate method in the context of Delphi studies. However, if it was chosen for Delphi studies, the decisions made by this test for some items (refer Table 4, e.g., item 29) might be wrong.

IV. Summary and Conclusion

A primary objective of a Delphi study is to obtain consensual and consistent responses from a group of experts' respondents between successive rounds. This stability issue was explored using the simulated dataset. The results from the four statistical tests were then evaluated. Findings from evaluating the simulated data were: (a). The McNemar change test could be robust even though the correction for continuity were not made, (b). The McNemar test is not as conservative as the Binominal test , and (c). The chi-square test for two independent samples could occasionally make wrong decisions when it is employed in Delphi studies.

The McNemar chi-square test is preferred to those parametric statistics (refer to the section of Delphi Research Method) that have been used in Delphi studies because:

(1). It provides a statistically significant test to decide whether experts have changed their mind between successive rounds. Some methods do not provide this utility and seem arbitrary and subjective for setting the stopping rules.

(2). It is especially suitable for the data collected from the Delphi study. The data collected from the Delphi study are often discrete rather than continuous. For the discrete data, computing the statistics such as the coefficient of variation (CV) values and the Pearson correlation coefficients seems inappropriate. In addition, when the samples are small (say 30) and not randomly drawn from the population of interest, the parametric statistics (e.g., F test) are not appropriate for this type of circumstance.

(3). It is rather easily computed. Along with the contingency table (refer to Table 1), the McNemar test is understandable to readers who do not necessarily have a strong background of statistics.

These are the appropriate steps in using the McNemar change test (Siegel & Castellan 1988).

(1). Cast the observed frequencies in a fourfold table as illustrated in Table 1.

(2). Determine the total number of changes (A+D). If the total number of changes is less than (or equal to) 10, use the binominal test.

(3). If the total frequency of changes exceeds 10, the McNemar is the right choice.

"Experts will make conjectures based upon rational judgment rather than merely guessing" (Weaver, 1971) is one of key principles that inherent in the Delphi studies. And, statistical techniques such as the McNemar change test and Binomial test provide a means for Delphi studies to ensure that the information gathered from experts are reliable.

REFERENCES

Blamphin, J. M., (1990). Applications of the focus group methodology for public relations research: A Delphi exploration of professional opinions. Unpublished master's thesis, University of Maryland, College Park, MD.

Dajami, J. S. (1979). Stability and agreement criteria for the termination of Delphi studies. Technological Forcasing and Social Change, 13, 83-90.

Dajani, J. S., Sincoff, M. Z. & Talley, W. K. (1979). Stability and agreement criteria for the termination of Delphi studies. Technological Forecasting and Social Change, 13, 83-90. Dalkey, N. C. (1967). *Delphi* (Report P-3704). Santa Monica, CA: Rand Corporation.

English, J. M. & Kernan, G. L. (1976). The prediction of air travel and aircraft technology to the Year 2000 using the Delphi method, Transport Research, 10, pp. 1-8.

Ferris, J. K. (1998). Leadership audit of university-based performing art center: A Delphi approach. Unpublished doctoral dissertation, The University of Maryland, College Park.

Franchak, S. J, Desy, J., & Norton, E. L. (1984). Involving business, industry, and labor: Guidelines for planning and evaluation vocational education programs. (Research and Development Series. No. 250). The Ohio State University at Columbus. The National Center for Research in Vocational Education.

Hendricks, S. M. (1995). Convergence on the roles and guidelines for designing, implementing and teaching a multicultural elementary art education curriculum. Unpublished doctoral dissertation. The University of Maryland, College Park.

Jolson, M. A. & Rossow, G. L. (1971). The Delphi process in marketing decision making. Journal of Market Research, 8, 443-448.

Linstone, H. A., & Turoff, M. (Eds.). (1975). The Delphi method: Techniques and applications. Massachusetts: Addison-Wesley Publishing Company. books, Henery Holt and Company, Inc.

Lucas, R. E. (1986). Evolving aesthetic criteria for computer generated art: A Delphi Study. Unpublished master's thesis, Ohio State University (ERIC Document Reproduction Service No. ED 286 462).

McNemar, Q. (1969). Psychological statistics. New York:Wiley.

Niero, M. & Robertson, A. (1996). Delphi estimates on clients' perceptions of family planning services. In M. Adler & E. Ziglio (Ed). Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health (pp. 213-237). Jessica Kingsley Publishers; Pennsylvania.

Preble, J. F. (1983). Public sector use of the Delphi technique. Technological Forecasting and Social Change, 23, 75-88.

Siegel, S. & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill Book Company.

Tafoya, W. L. (1986). A Delphi forecast of the future of law enforcement. Unpublished doctoral dissertation. The University of Maryland, College Park.

Toohey, J.V. & Shireffs, J. H. (1980). Health Education, 11, 15-17.

Wang, T. C., (1992). Present and potential instructional use of computers in Art: A Delphi study. Unpublished doctoral dissertation, The University of Maryland, College Park.

Weaver, W. T. (1971). The Delphi forecasting method. Bloomington, In: Phi Delta Kappan, 52 (5), 267-273, pp 268.

Yang, Y. N. (2000, April). Convergence on the guidelines for designing a web-based Art teacher education curriculum: A Delphi study. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: info@ericfac.piccard.csc.com
WWW: http://ericfacility.org